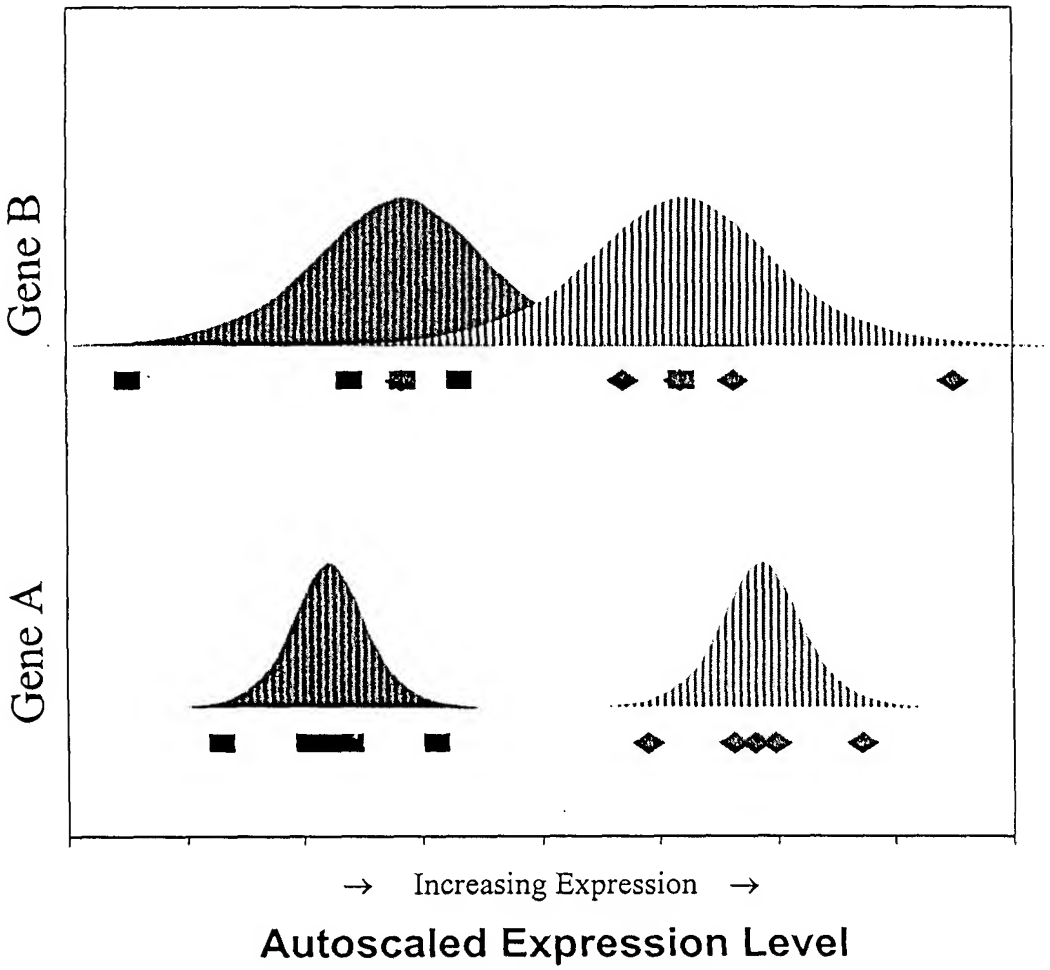


Figure 1: A comparison of two idealized gene expression distributions. Gene A represents a good discriminator, because the variance within each group is significantly smaller than the between-group variance. Gene B, on the other hand, has significant overlap between the two groups relative to their variances. To capture this distinction for a gene i that may be classified in j classes, we calculate the total variance (V^t_i) and compare it to the within cluster variance (V^c_i) to determine the Wilks' lambda score (Λ_i): lower scores mean better discrimination.

$$V_i^t = (x_i - \bar{x}_i)^T (x_i - \bar{x}_i)$$

$$V_i^c = \sum_{j=1}^{n} V_i^{cj} = \sum_{j=1}^{n} (x_i^j - 1\bar{x}_i^j)^T (x_i^j - 1\bar{x}_i^j)$$

$$\Lambda_i = \frac{V_i^c}{V_i^t}$$

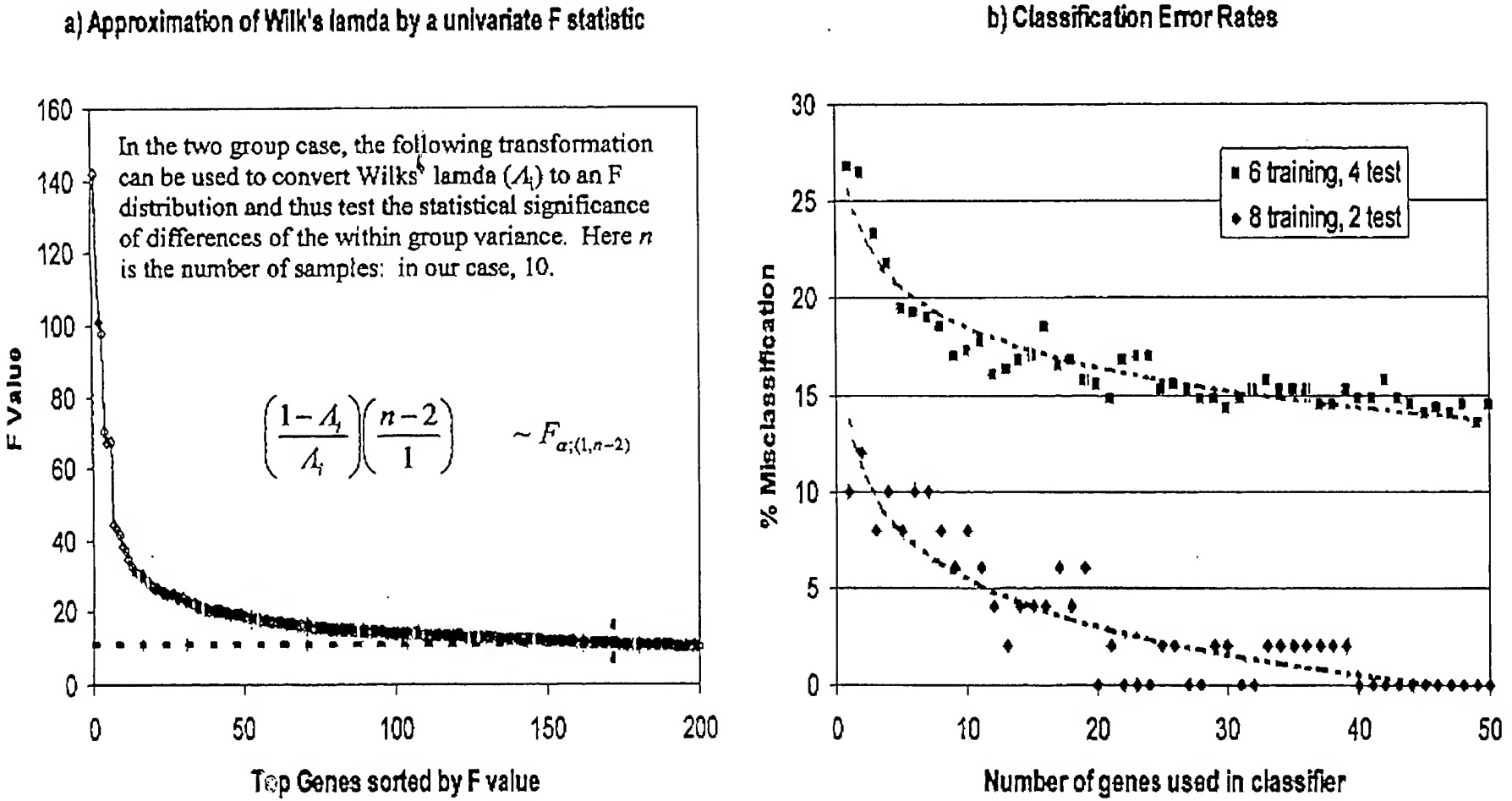


Figure 2: a) The top 200 discriminating genes rank ordered by the F statistic (equation in the figure). The top 171 genes give an F value above the $\alpha = 0.99$ confidence interval ($F_{\alpha,(1,8)} = 11.26$), providing a conservative estimate of the genes characteristic of the cancerous state. b) Cross-validation results of classification through the CDA analysis to determine discriminatory genes. In each case, the best discriminatory genes were determined independently of the other cases, so each classification result arises from a different sub-set of the genes. Perfect classification requires about 45 genes when eight samples are included in the training set.

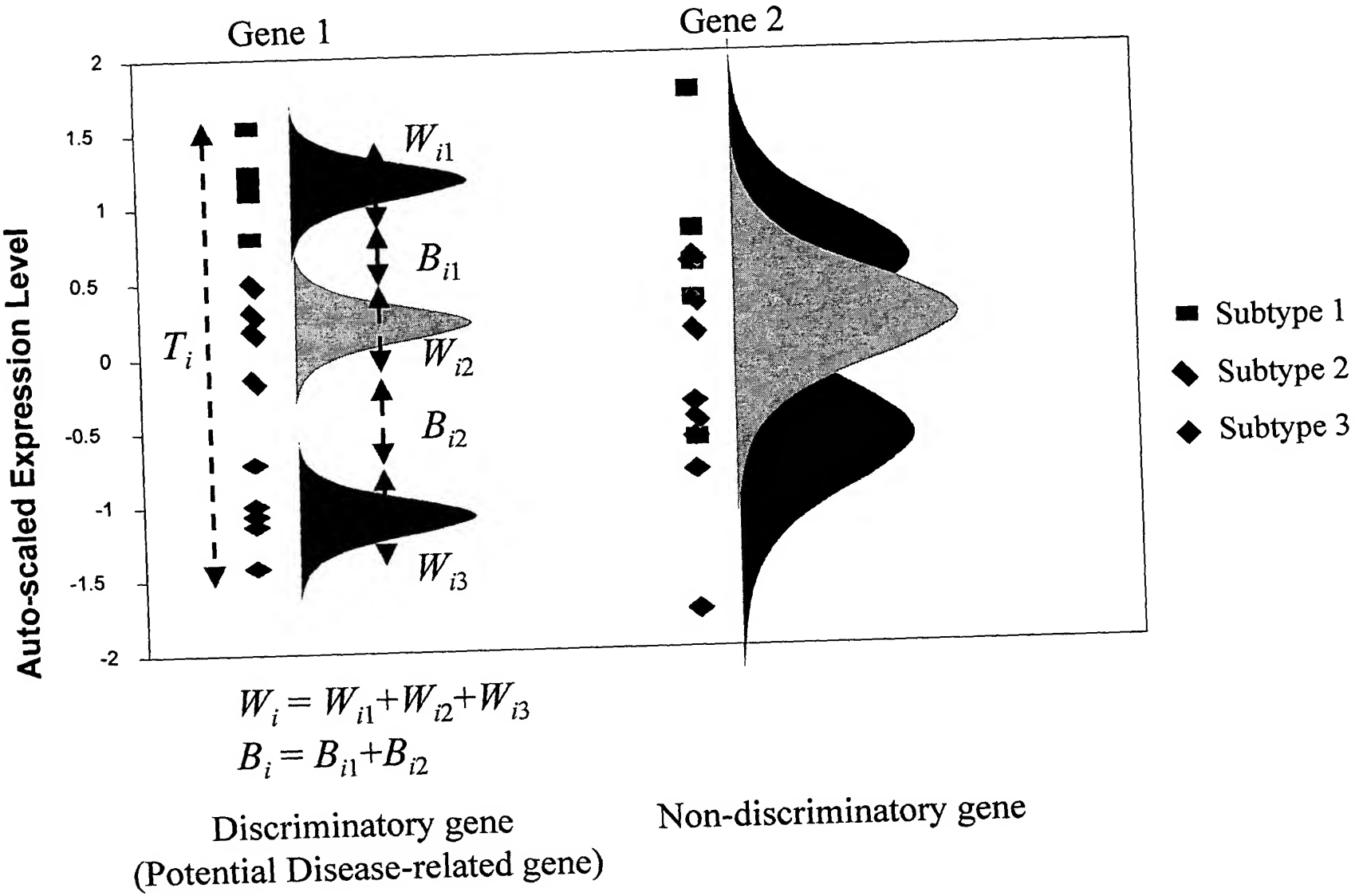


Figure 3. Discriminatory genes (potential disease related genes) and Non-discriminatory genes.

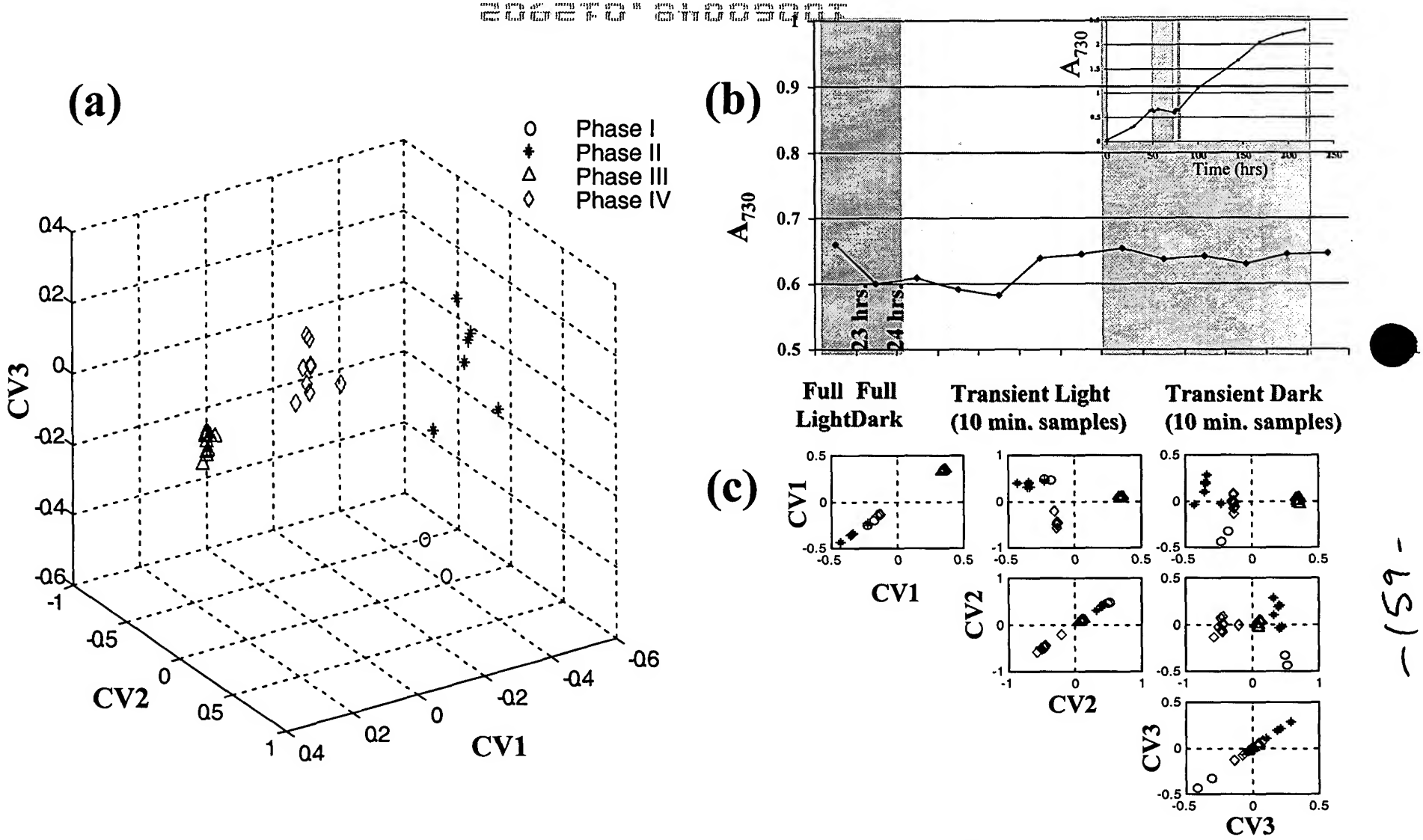


Figure 4: a)Projection of the expression phenotypes of cultures of *Synechocystis* sp. PCC 6803 to a CDA-defined space. This photosynthetic bacterium was grown under conditions shown in b) and the expression levels of 87 genes were measured by a DNA microarray at 29 time points spanning the entire course of the experiment. Of the 87 genes, 27 were identified as most discriminating of the four classes defined by the four different light conditions and their expression levels were projected to the CDA-defined space. It can be seen that the four phenotypic classes are clearly identified in the 3-dimensional CDA projection space. c) The first CV shows the largest discrimination power separating all the groups, discriminating clearly Phase III from the others. The second CV separates Phase IV from Phases I and II, while the third CV is necessary to separate Phase I from II.

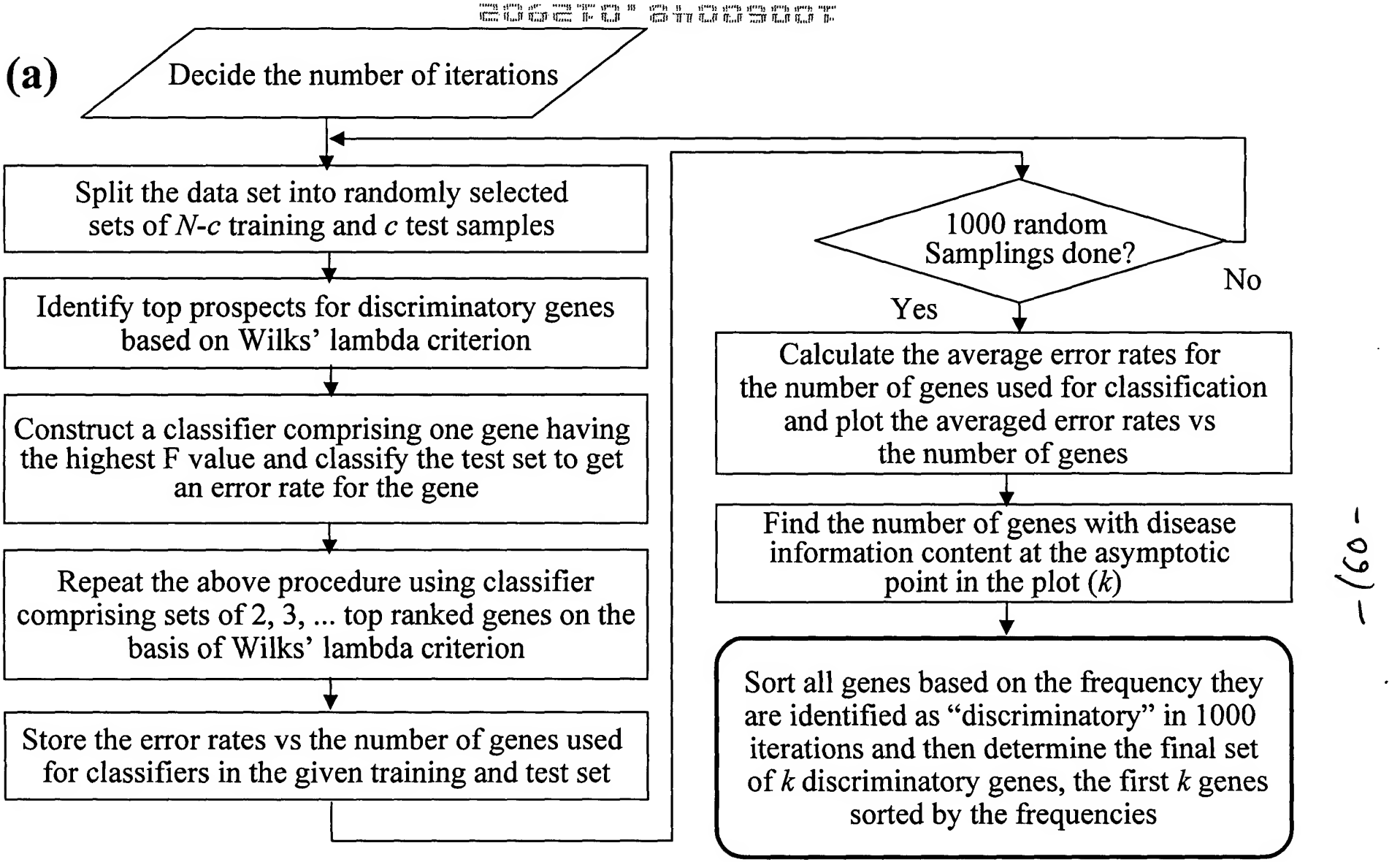


Figure 5a. Leave one out cross-validation (LOOCV) algorithm, where *N* is the total number of samples and *c* is the number of classes, so that one sample from each class is included in the test.

(b)

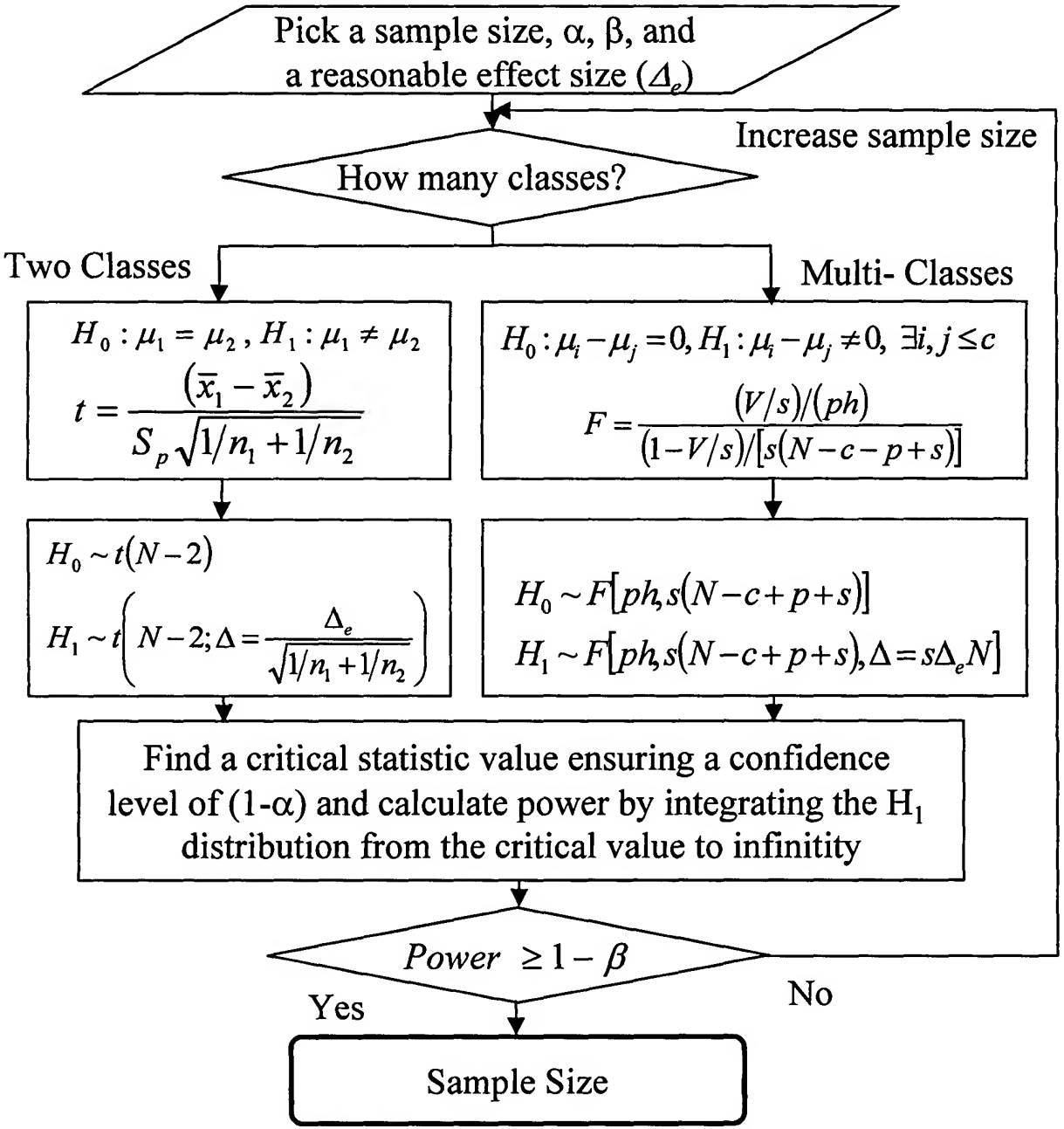


Figure 5b. Power analysis algorithm for determination of the minimum sample size.

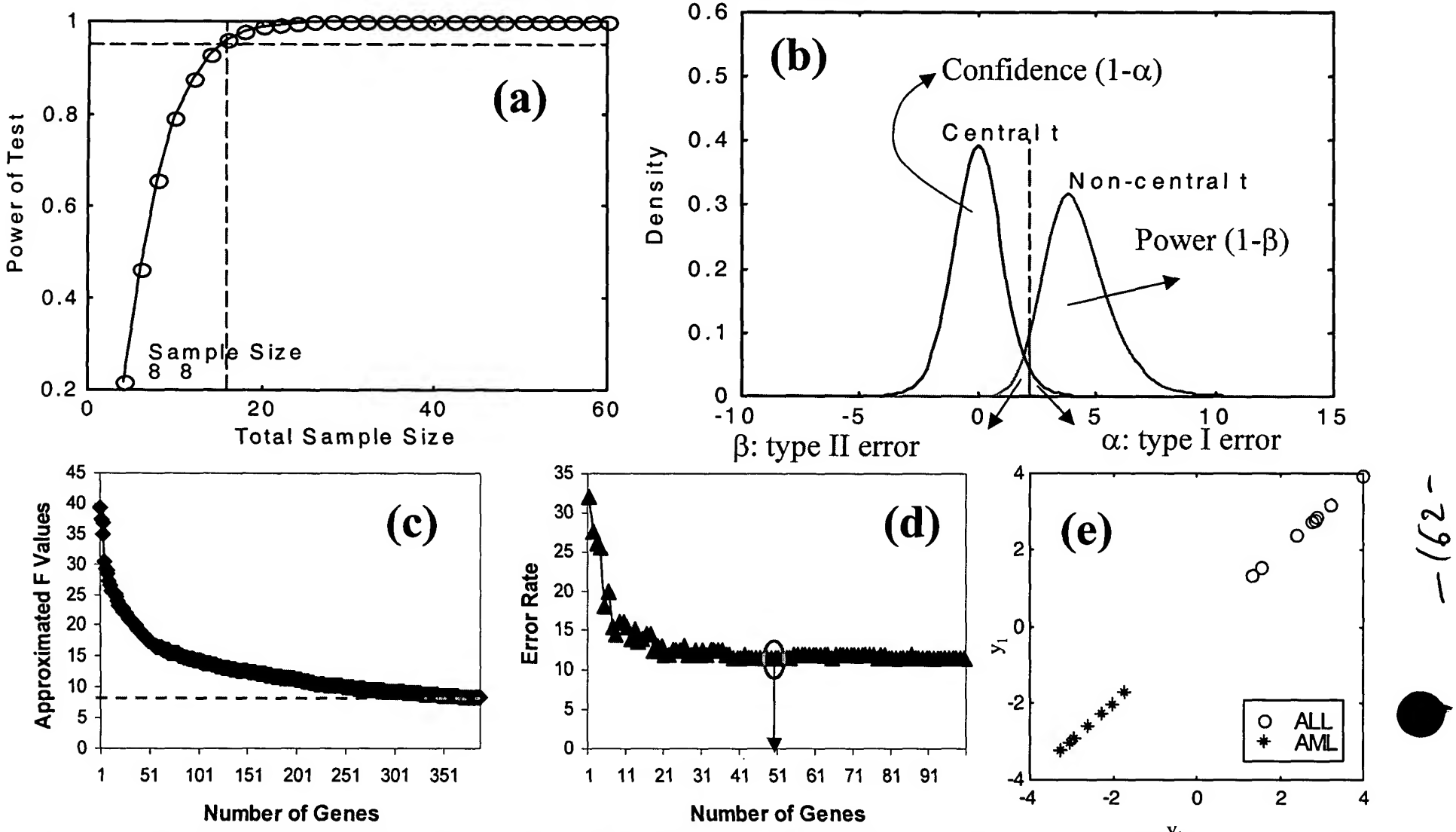


Figure 6. determination of minimum sample size for two-class (ALL, AML) distinction, selection of discriminatory genes with the estimated sample sizes of two classes, and FDA projection. (a) Power plot versus sample size showing how to determine the sample size required for two class distinction (8 from each class). (b) The distributions of H_0 and H_1 for the determined sample size. (c) Univariate F statistic values of the initial 388 discriminatory genes with a threshold ($F_{0.01(1,18)} = 8.2854$) in randomly selected 8 ALL and 8 AML samples out of the entire data set. (d) Leave-one-out cross-validation applied to estimate the classification error rates and then to select the 50 most discriminatory genes with the same samples. (e) Separation of the 8 ALL and 8 AML samples in the two-dimensional FDA projection space defined discriminant axes of the 50 discriminatory genes.

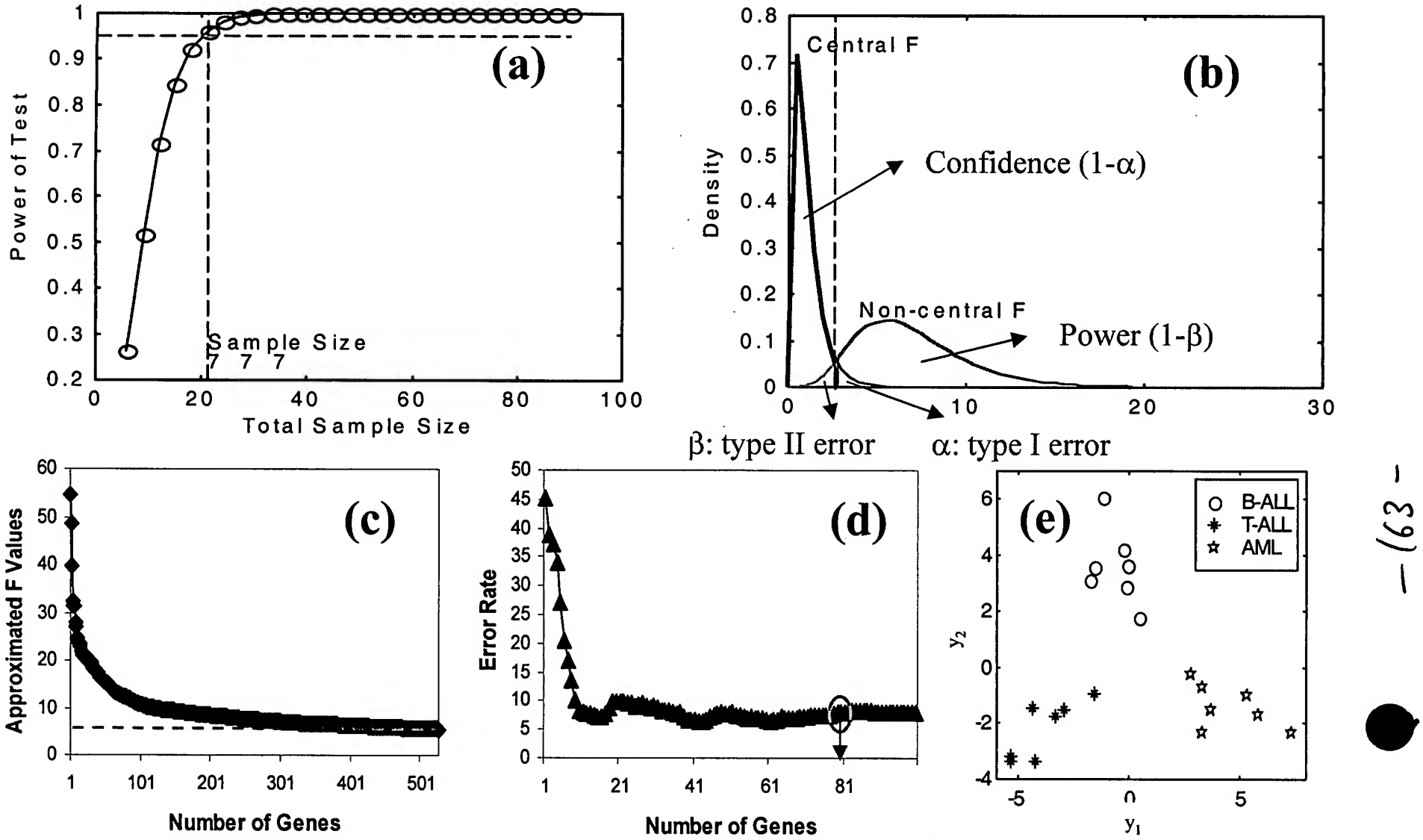


Figure 7. determination of minimum sample size for the three-class (B-ALL, T-ALL, AML) distinction, selection of discriminatory genes with the estimated sample sizes of three classes, and FDA projection. (a) Power plot versus sample size showing how to determine the minimum sample size (7 from each class). (b) The distributions of H_0 and H_1 for the determined sample size. (c) Univariate F statistic values of the initial 527 discriminatory genes with a threshold ($F_{0.01(2,26)} = 5.5263$) in randomly selected 7 B-ALL, 7 T-ALL and 7 AML samples out of the entire data set. (d) Leave-one-out cross-validation applied to estimate the classification error rates and then to select the 80 most discriminatory genes with the same samples. (e) Separation of the 7 B-ALL, 7 T-ALL and 7 AML samples in the two-dimensional FDA projection space defined discriminant axes of the discriminatory 80 genes.

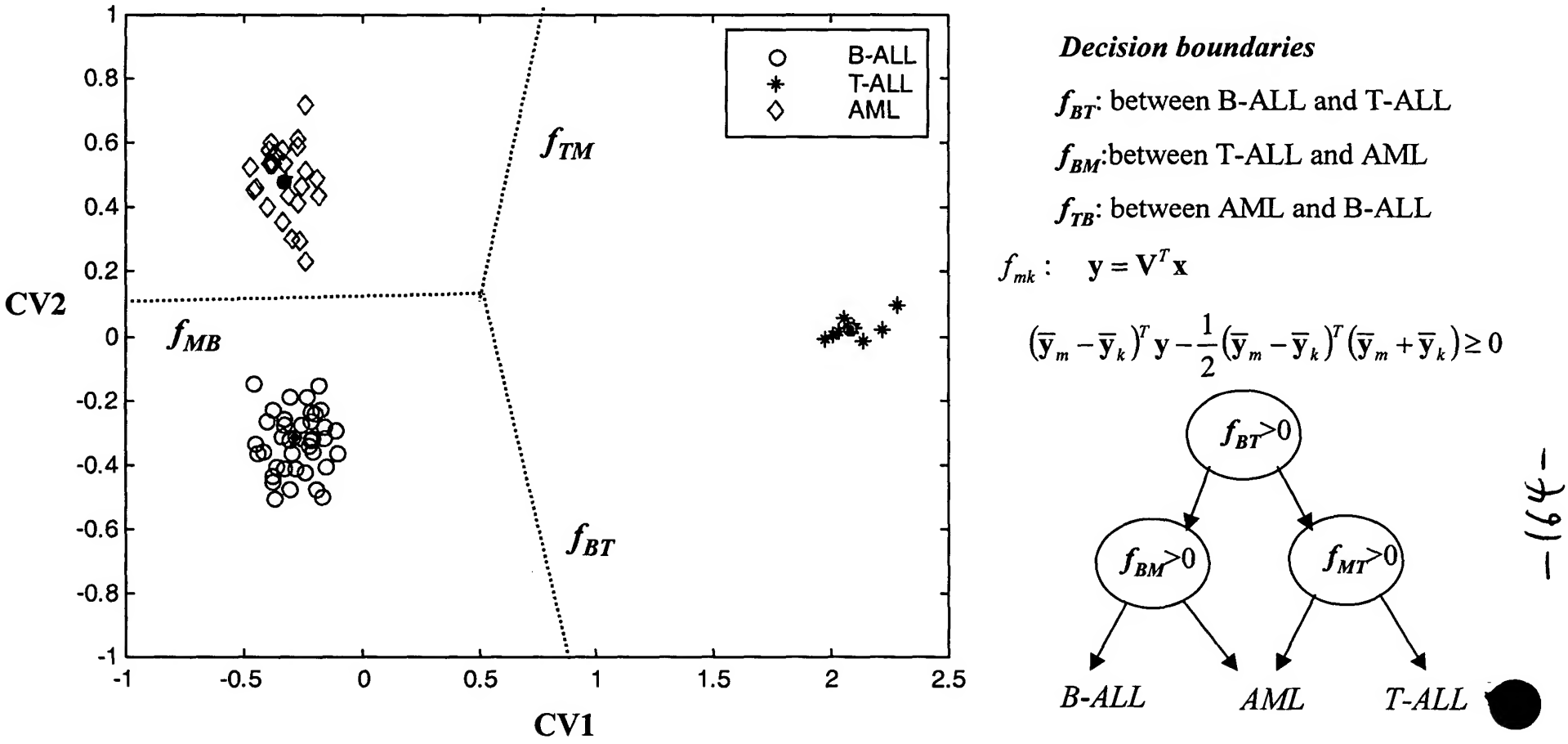


Figure 8: Physiological state domains are defined by a series of decision boundaries for the projected values in the reduced CV plane. Geometrical information can be inferred from this projection: for example, the covariance structure in B-ALL group is different from those in other groups. The decision boundary can be defined as the line perpendicular to the line joining two adjacent group means (each group mean is represented by a solid green circle). A classification tree can then be built for class prediction using those decision boundaries, as shown on the bottom right.

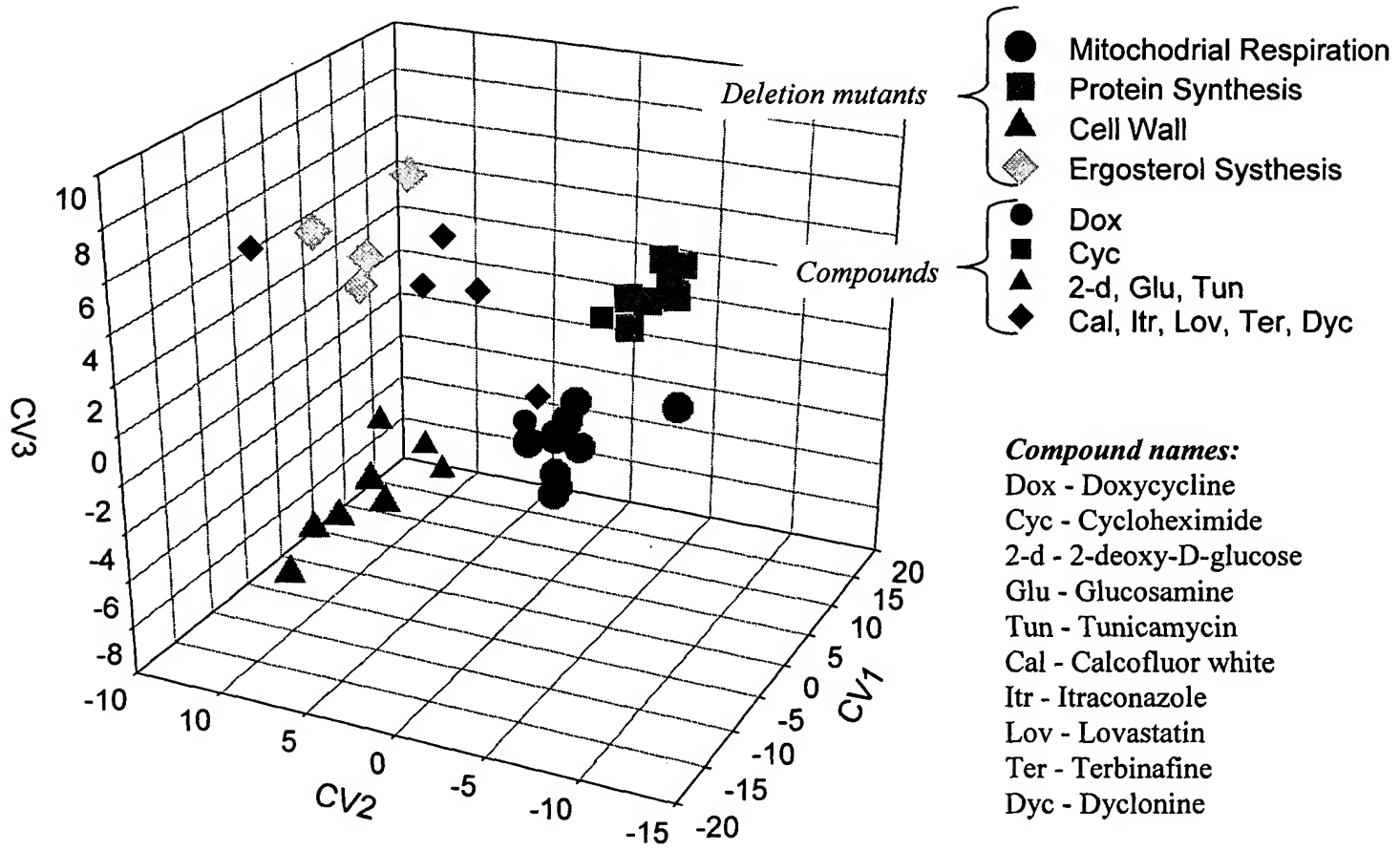
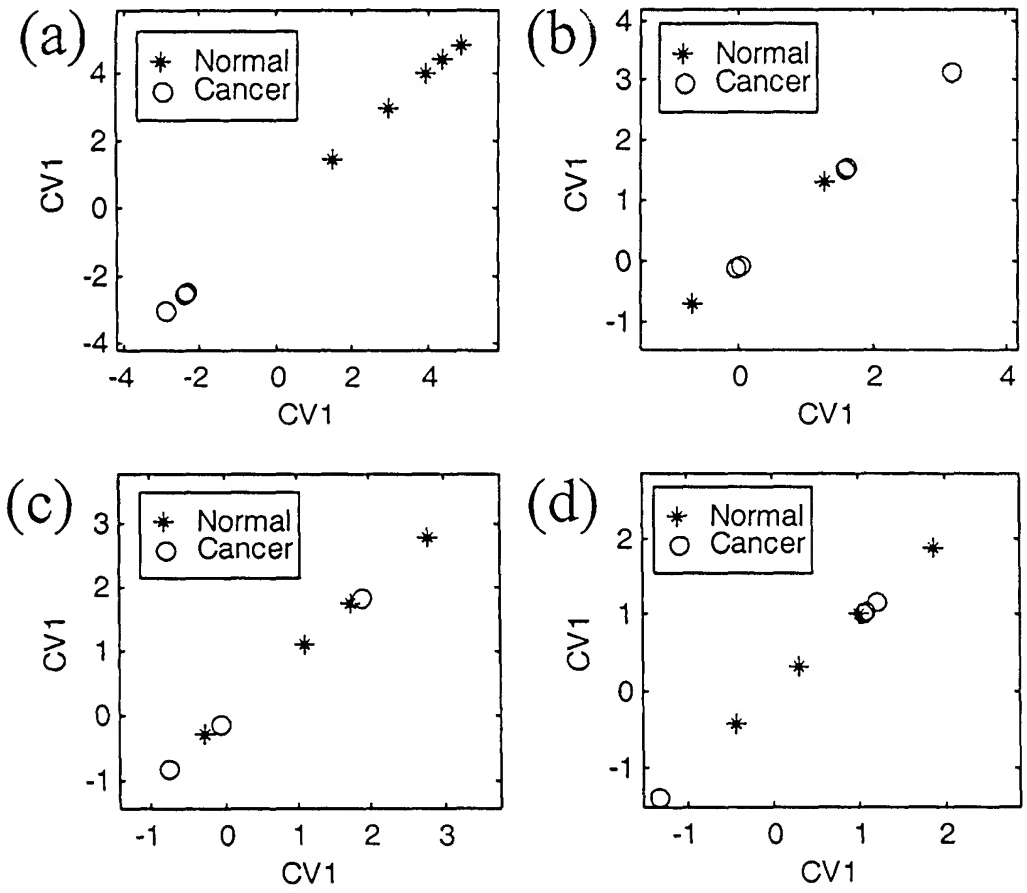


Figure 9: CDA projection of 27 yeast deletion mutant expression phenotype experiments grouped by the functionality of the eliminated gene. Four groups of related mutants have been distinguished using three CVs by projecting the expression levels of 200 of the most discriminating genes. The expression phenotypes obtained from the application of 10 chemical compounds to the wild-type yeast cultures are also projected into the CDA space defined by the mutants. The proximity in CDA space of these projections to those of the expression phenotype of the deletion mutant groups helps characterize the action of the compound on cell physiology. Note that one compound experiment (Cal) which appears incorrectly classified is actually in the center of the 3-D diagram, and not clearly associated with any of the groups shown. The classification suggested by the proximity of the projected phenotypes to the deletion mutants groups agrees with classification provided by Huges *et al.* (2000).



Canonical Discriminant Analysis (CDA) filters out the most relevant discriminatory information from the data space by projecting the samples into a reduced space. This reduced space is defined by a set of Canonical Variables (CV), which are a linear combination of the discriminatory genes.

The following optimization problem is solved to compute each CV, denoted by v.

$$\underset{\mathbf{v}}{Max}\frac{\mathbf{v}^{T}\mathbf{W}^{-1}\mathbf{B}\mathbf{v}}{\mathbf{v}^{T}\mathbf{v}}$$

$$\mathbf{W}=\sum_{k=1}^{c}\left(\mathbf{X}_k-\bar{\mathbf{x}}_k\mathbf{1}^{T}\right)\left(\mathbf{X}_k-\bar{\mathbf{x}}_k\mathbf{1}^{T}\right)^{T}$$

$$\mathbf{T}=\sum_{k=1}^{c}\left(\mathbf{X}-\bar{\mathbf{x}}\mathbf{1}^{T}\right)\left(\mathbf{X}-\bar{\mathbf{x}}\mathbf{1}^{T}\right)^{T}\qquad \mathbf{B}=\mathbf{T}-\mathbf{W}$$

Here **W** represents the within-group variance, **B** represents the between-group variance, and **T** represents the total variance. The solution is the matrix of eigenvectors (**V**) of $\mathbf{W}^{-1}\mathbf{B}$ whose corresponding eigenvalues (Λ) represent the relative discriminatory power of each CV.

$$\mathbf{W}^{-1}\mathbf{B}=\mathbf{V}\Lambda\mathbf{V}^{T}$$

Figure 10: CDA results using 45 genes. The projected value (CDA scores) into CV were calculated using linear combination of individual gene expressions, $CV=v_1g_1+v_2g_2+\ldots+v_{45}g_{45}$. (a) Top 45 discriminatory genes, (b) 2000th-2045th genes, (c) 4000th-4045th genes, (d) 6000th-6045th genes. The overlap and variation in the groups increases when genes are chosen poorly.

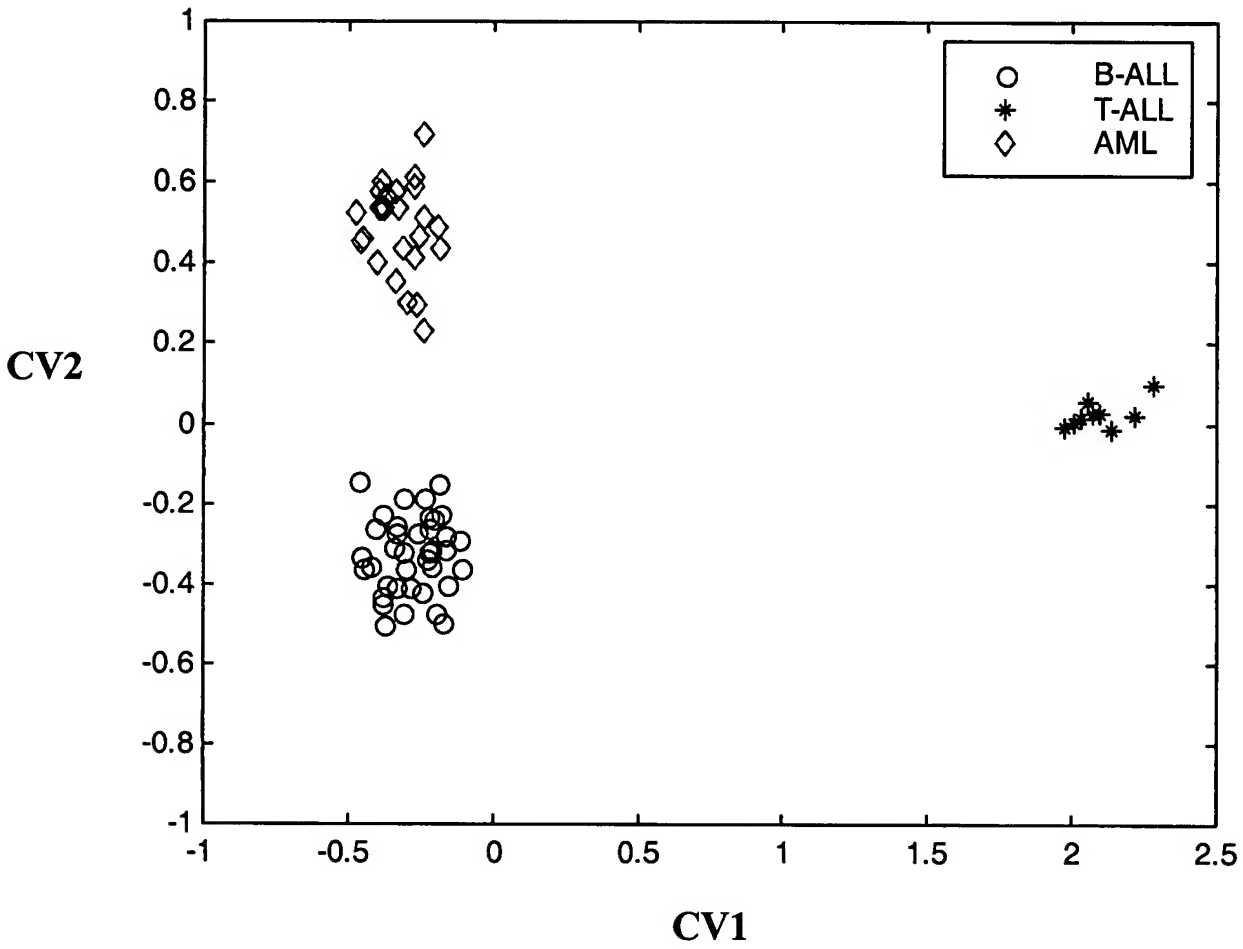


Figure 11: CDA projection of expression data obtained from patients with B-ALL, T-ALL, and AML. Projection of 30 discriminatory genes allow CDA to clearly separate the three classes of leukemia expression phenotype in a two dimensional CV plane. The first CV distinguishes the B-ALL group from T-ALL and AML. The second CV separates T-ALL group from AML to complete the group separation.

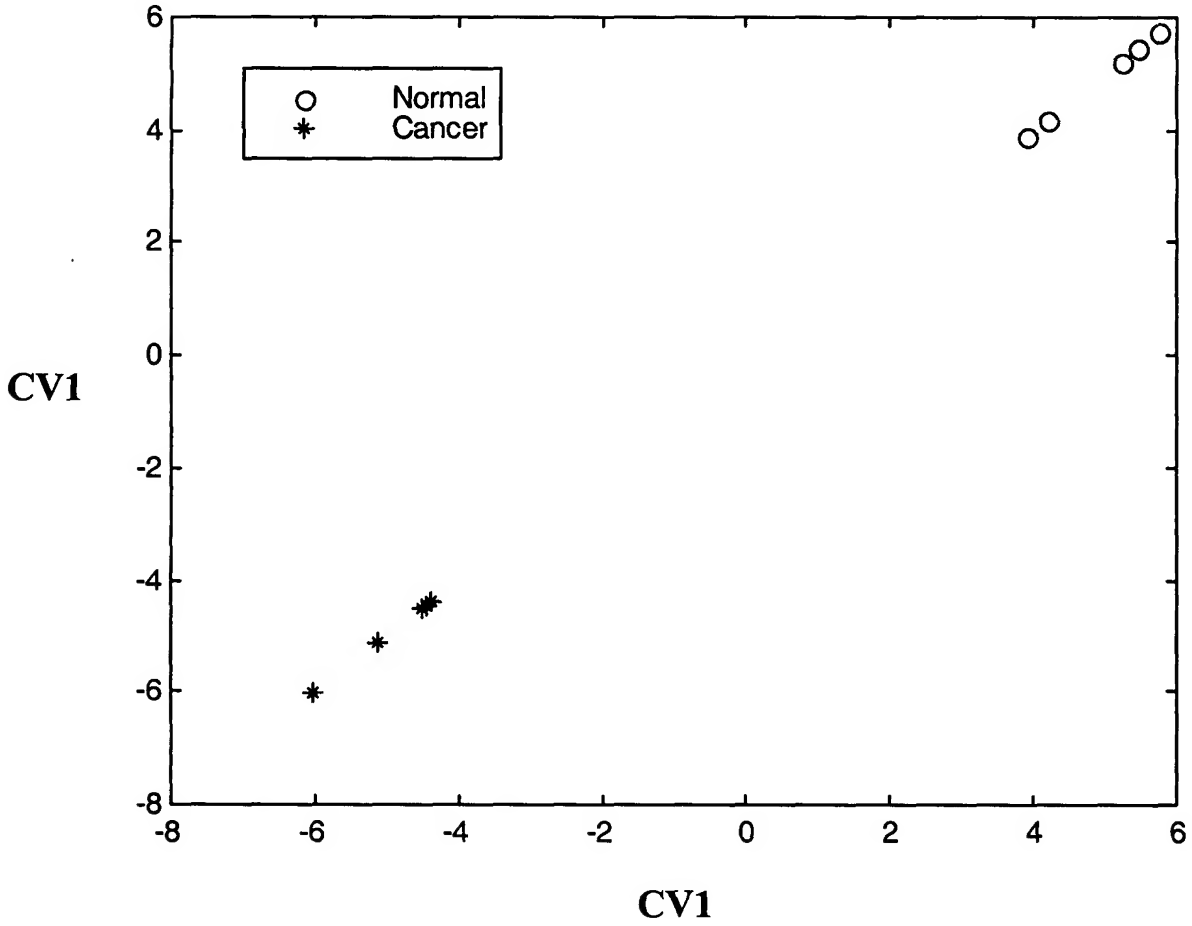


Figure 12: CDA projection of the expression phenotypes comprising 7070 genes measured in samples obtained from healthy individuals (5 samples) and patients with oral epithelium cancer (5 samples). 35 discriminatory genes out of 7070 total genes allow CDA to clearly separate the two groups in one dimensional CV line. The CV value distinguishes normal tissues from cancerous tissues.

Figure 13

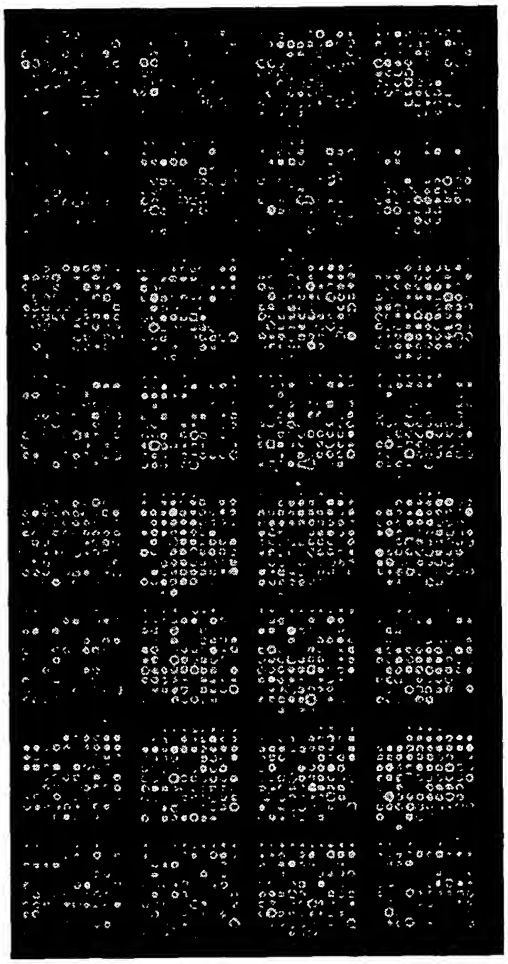
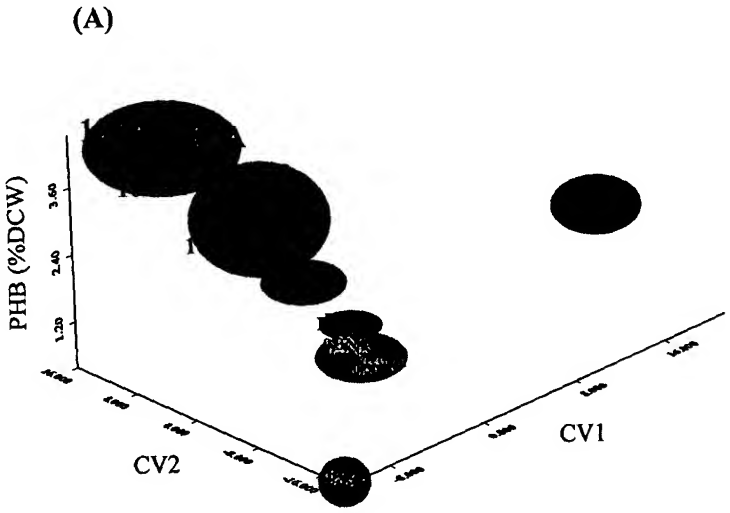
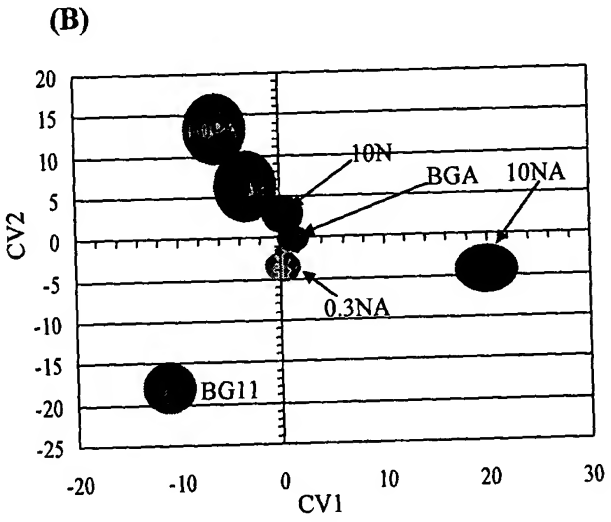


Figure 14


(A)
PHB (%DCW)
CV2
CV1


(B)
CV2
CV1
10N
BGA
10NA
0.3NA
BG11

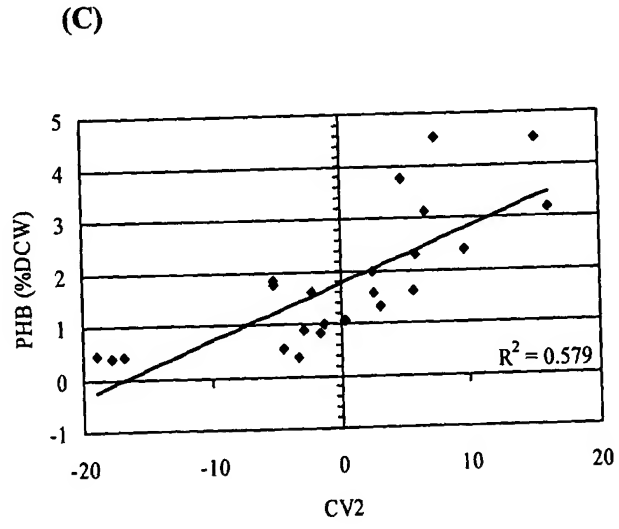
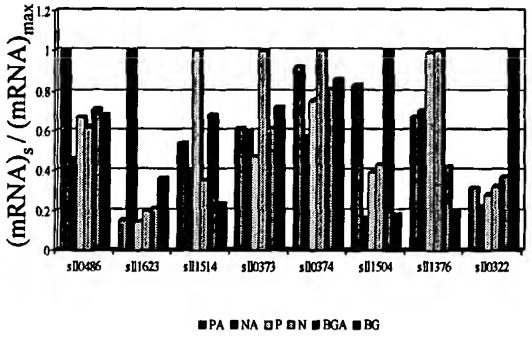
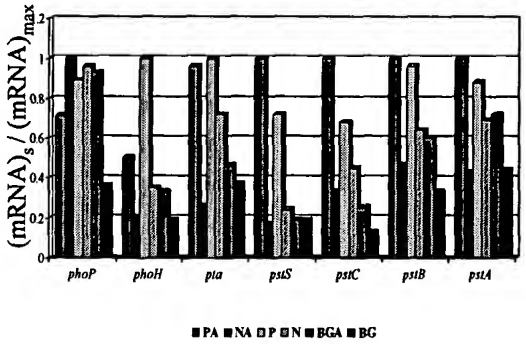

(C)
PHB (%DCW)
CV2
R2 = 0.579

Figure 15

(A)



(B)



(C)

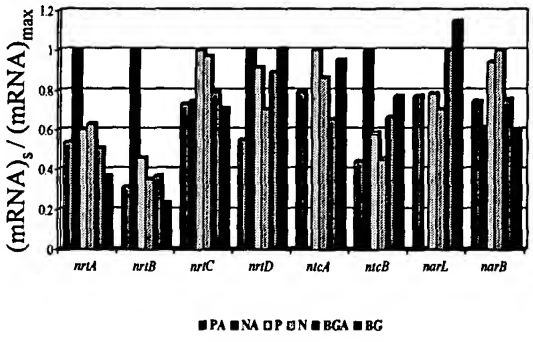
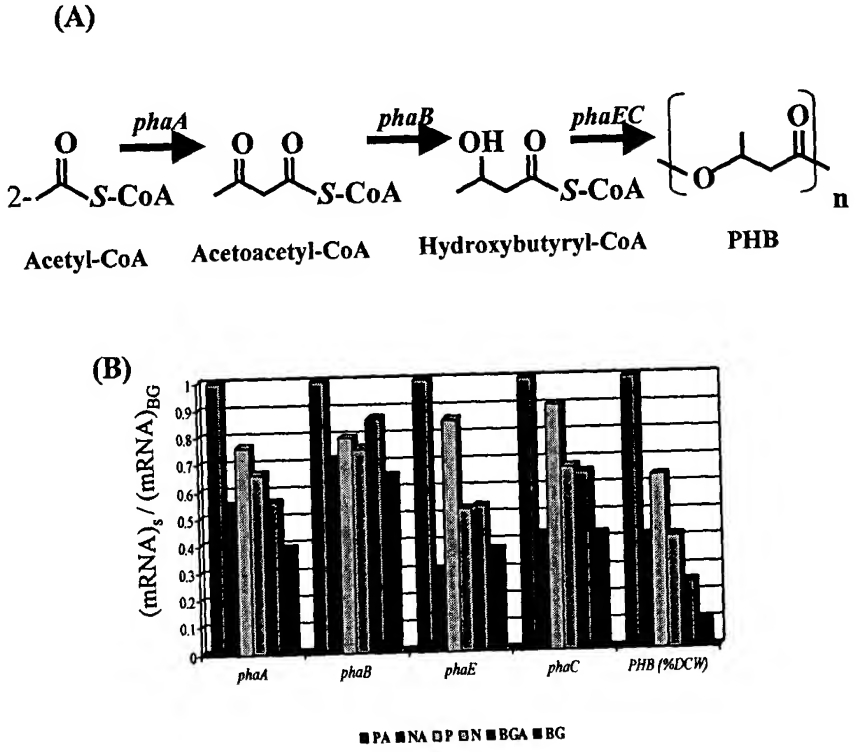


Figure 16


(A)
phaA
phaB
phaEC
2-
S-CoA
OH
n
Acetyl-CoA
Acetoacetyl-CoA
Hydroxybutyryl-CoA
PHB
(B)
(mRNA)s / (mRNA)BG
phaA
phaB
phaE
phaC
PHB (%DCW)
PA NA P N BGA BG